# CSCI 699: Trustworthy ML (from an optimization lens)

Vatsal Sharan

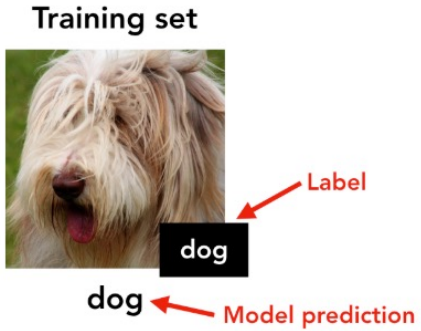
Fall 2025

Lecture 5, Sep 24

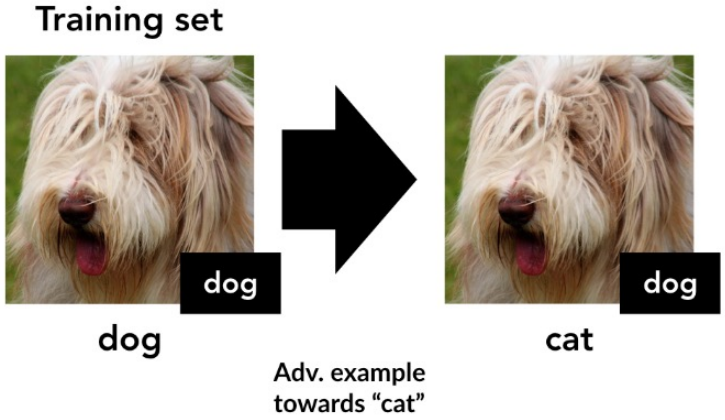
USC University of Southern California

# Recap

# Robust vs non-robust features




Consider an image, classified correctly




Perturb image, to get an adversarial example. Image is now misclassified

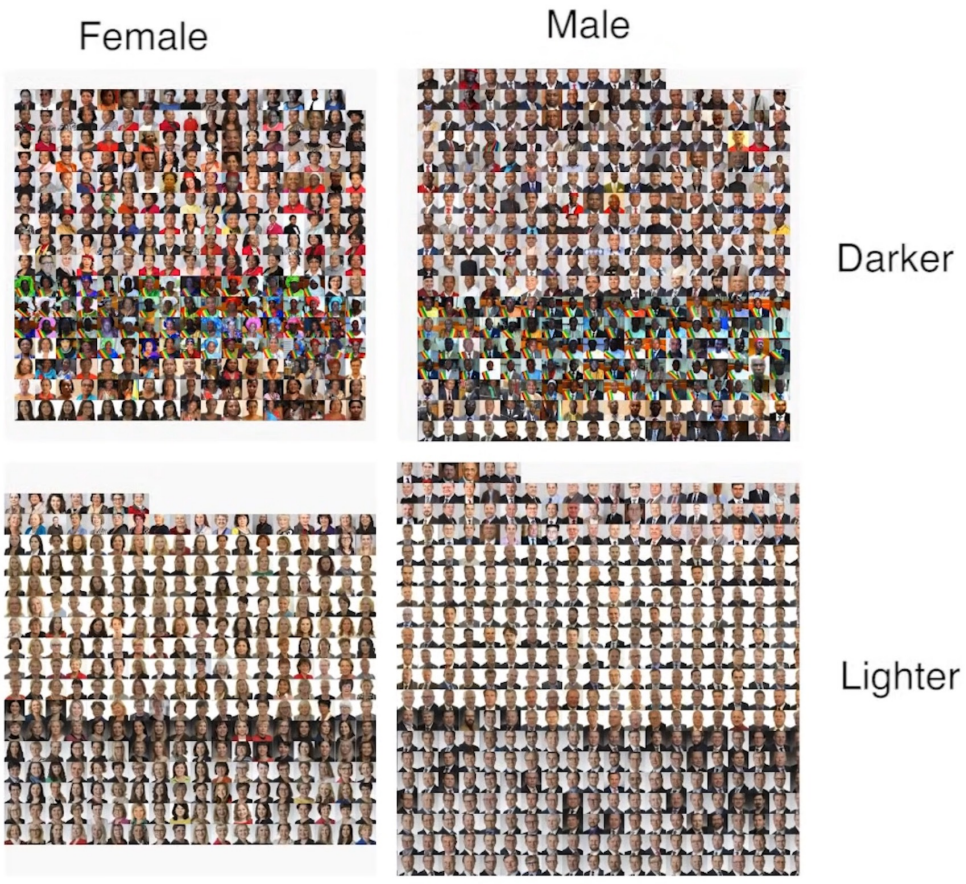

Label the adversarial image with the incorrect label

- Suppose we take CIFAR10 and a model trained on CIFAR10, replace each image by its adversarial example for some class, and "relabel" the image with this wrong class.
- Now train a model on this new CIFAR10, and then evaluate on the normal CIFAR10 test set. How much accuracy do we expect?

- Model gets highly non-trivial accuracy! ($\approx 45\%$ on 10 class classification)

Fig. from *Adversarial Examples Are Not Bugs, They Are Features, Ilyas et al. '19. See Blog post at* [https://gradientscience.org/adv/](https://gradientscience.org/adv/)

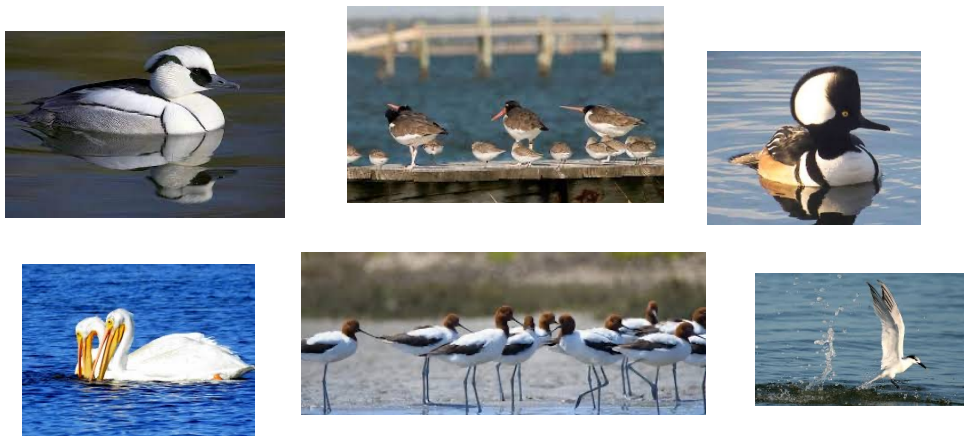
# Poor performance on subgroups: Gendershades




http://gendershades.org/

# ML models can latch onto spurious features to make predictions

Most images of waterbirds are in water,
and landbirds are on land



Waterbirds vs. Landbirds

# Distribution shifts: Setup

**What if we get training samples from $\mathcal{D}$, but test samples from $\mathcal{D}'$?**

$\mathcal{D}'$ can differ from $\mathcal{D}$ in two of these ways:

- Let $p(x)$ and $p'(x)$ be marginals of $x$ under $\mathcal{D}$ and $\mathcal{D}'$. Then $p'(x)$ may be different from $p(x)$. This is known as a *covariate shift*, only the covariates $x$ have changed.
- The conditional distribution $\Pr_{\mathcal{D}}[y|x]$ may be different from $\Pr_{\mathcal{D}'}[y|x]$. This is known as a *concept shift*. Here the ground-truth itself has changed.

For covariate shifts, we can loosely split them into two kinds of shifts the community thinks about:

- When $p(x)$ and $p'(x)$ are collected from independent and potentially different data collection processes, for example data from two different hospital systems. We saw this in class presentations last week.
- When $p'(x)$ can be regarded as a reweighting of $p(x)$, for example considering the group of "darker skinned females" for facial recognition, or "images of waterbirds on land background" for the landbirds/waterbirds task. This is also known as *subgroup robustness*.

# Distributionally robust optimization for subgroup robustness

In usual supervised ML we care about finding some predictor $f^*$ such that

$$f^* := \arg\min_{f \in \mathcal{F}} \Big\{ \mathbb{E}_{(x,y)\sim\mathcal{D}}[\, \ell(f(x), y)\, ] \Big\}.$$

Suppose we have a set of groups $g \in \mathcal{G}$, each of which defines some distribution $\mathcal{D}_g$ (which could be a re-weighting of $\mathcal{D}$ with respect to the marginal of $x$). Then we can define the distributionally robust formulation of ML as:

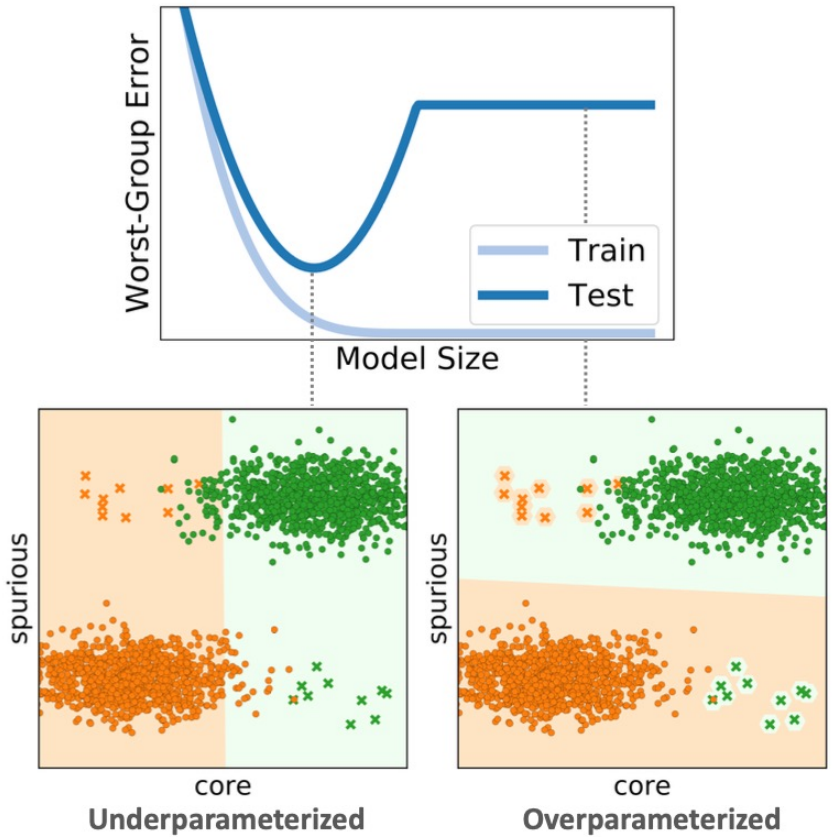
$$f^*_{\text{DRO}} := \arg\min_{f \in \mathcal{F}} \Big\{ \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y)\sim D_g}[\, \ell(f(x), y)\, ] \Big\}.$$

As is usual in supervised ML, we do not actually have access to the distribution $D_g$, but work with empirical samples.

$$\hat{f}^*_{\text{DRO}} := \arg\min_{f \in \mathcal{F}} \Big\{ \max_{g \in \mathcal{G}} \frac{1}{|\#\text{samples from group } g|} \sum_{(x,y) \in \text{ group } g} \ell(f(x), y) \Big\}.$$

Also see *Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization, Sagawa et al. '20*

# Worst-group generalization, and importance of regularization




Overparameterized models use the signal from majority group (so relying on the spurious feature here), and “memorize” the minority group samples

Need to add regularization to get generalization on minority group

Fig from *An Investigation of Why Overparameterization Exacerbates Spurious Correlations, Sagawa et al. ‘20*

# The ML loop

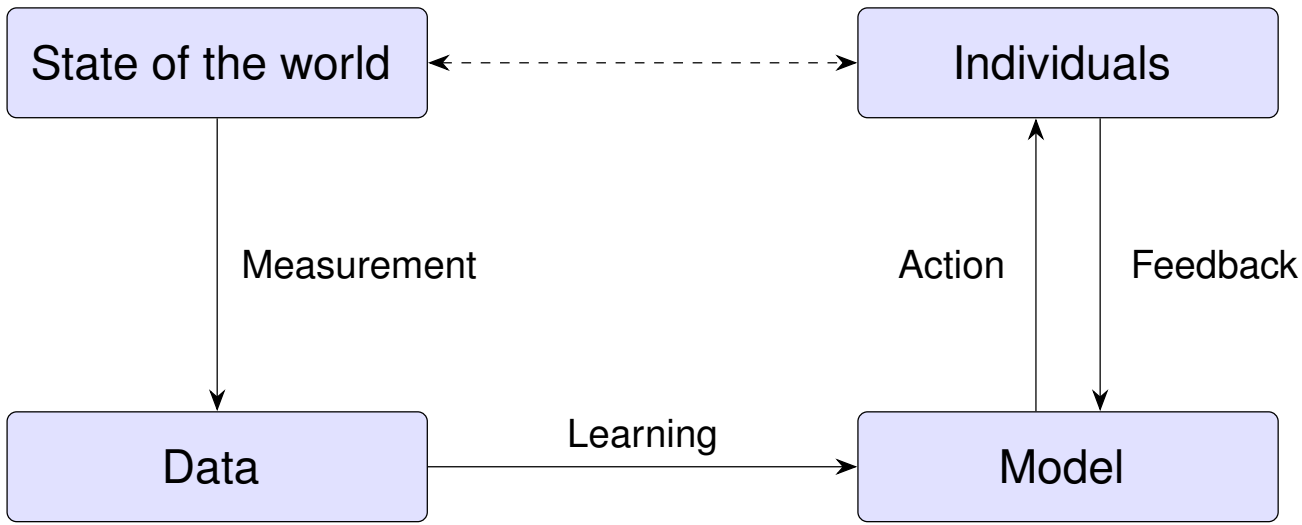
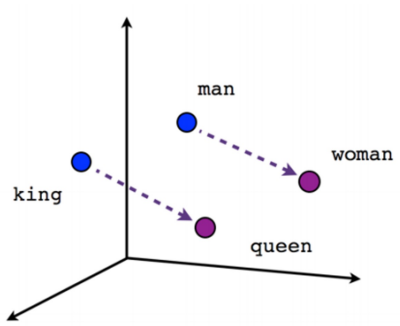



Fig. based on the book *Fairness And ML: Limitations and Opportunities*

# Bias in representation: Word embeddings




Male-Female

Word analogy questions:

man: woman :: king : ??

$$\overrightarrow{\text{woman}} - \overrightarrow{\text{man}} \approx \overrightarrow{\text{queen}} - \overrightarrow{\text{king}}$$

$$\overrightarrow{\text{woman}} - \overrightarrow{\text{man}} \approx \overrightarrow{\text{homemaker}} - \overrightarrow{\text{computer programmer}}$$

**Gender stereotype *she-he* analogies.**

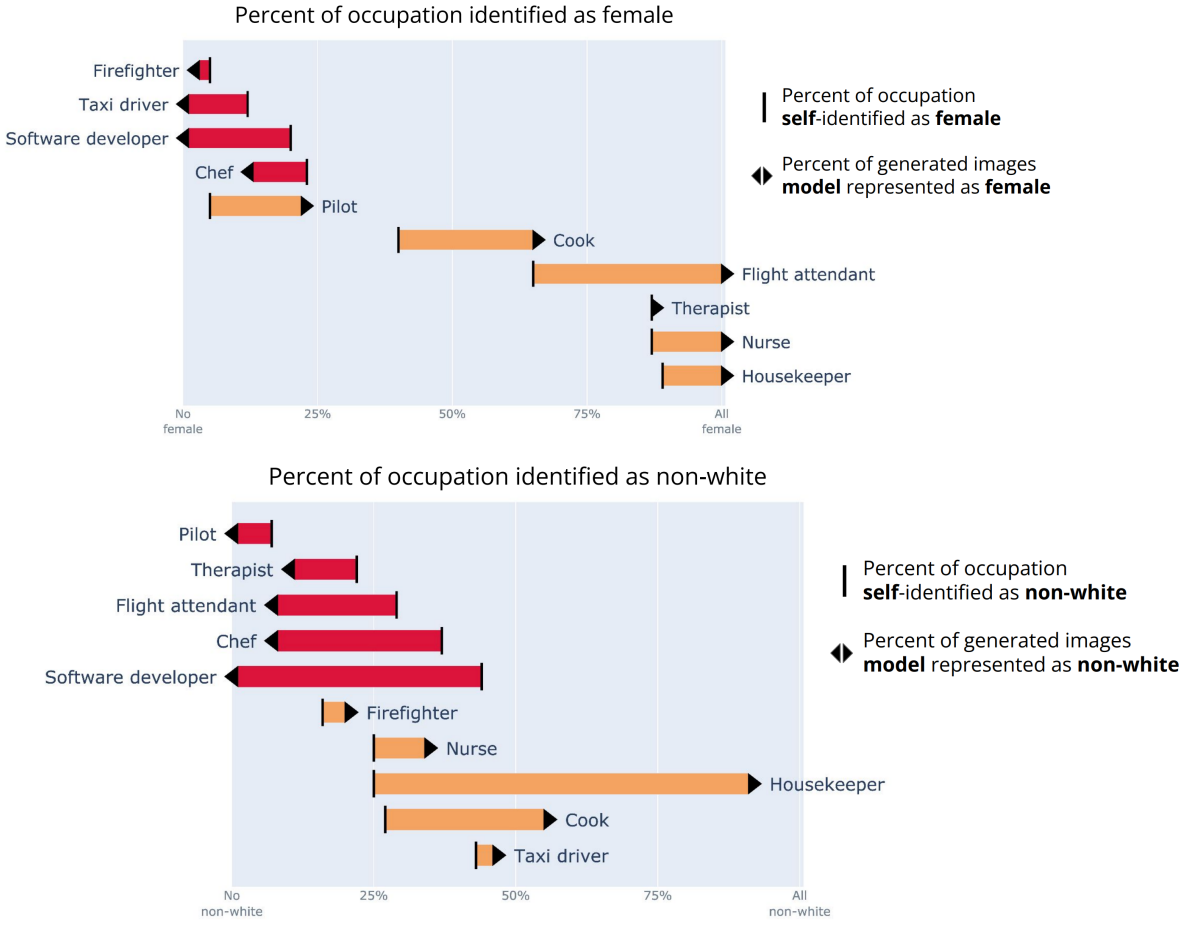
| | | |
|---|---|---|
| sewing-carpentry | register-nurse-physician | housewife-shopkeeper |
| nurse-surgeon | interior designer-architect | softball-baseball |
| blond-burly | feminism-conservatism | cosmetics-pharmaceuticals |
| giggle-chuckle | vocalist-guitarist | petite-lanky |
| sassy-snappy | diva-superstar | charming-affable |
| volleyball-football | cupcakes-pizzas | hairdresser-barber |

**Gender appropriate *she-he* analogies.**

| | | |
|---|---|---|
| queen-king | sister-brother | mother-father |
| waitress-waiter | ovarian cancer-prostate cancer | convent-monastery |

From *Man is to Computer Programmer as Woman is to Homemaker?* Debiasing Word Embeddings, Bolukbasi et al. '16

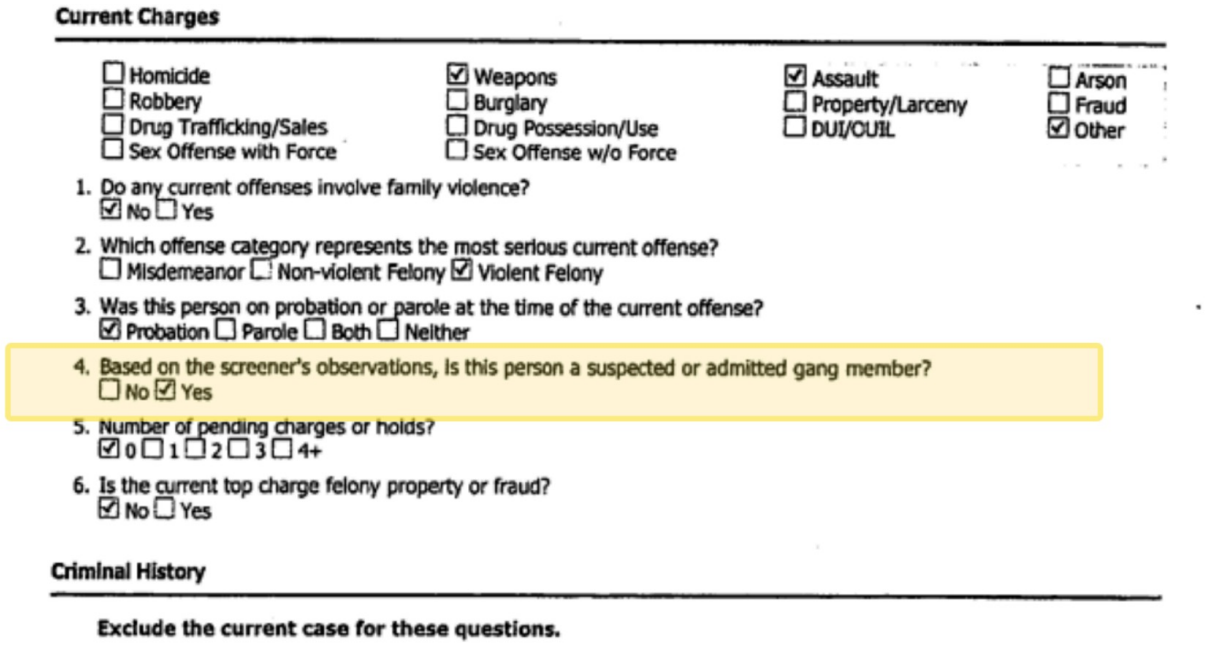
# Text to image models can amplify existing biases




*Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale*, Bianchi et al., 2023
For more discussion, see *A Systematic Study of Bias Amplification,* Hall et al., 2022

# Bias in predictions: The COMPAS software

- COMPAS is a proprietary software used by many judicial systems to determine the risk that someone arrested for a crime again commits a crime in the future
- Used for decisions such as for deciding bail



**Current Charges**

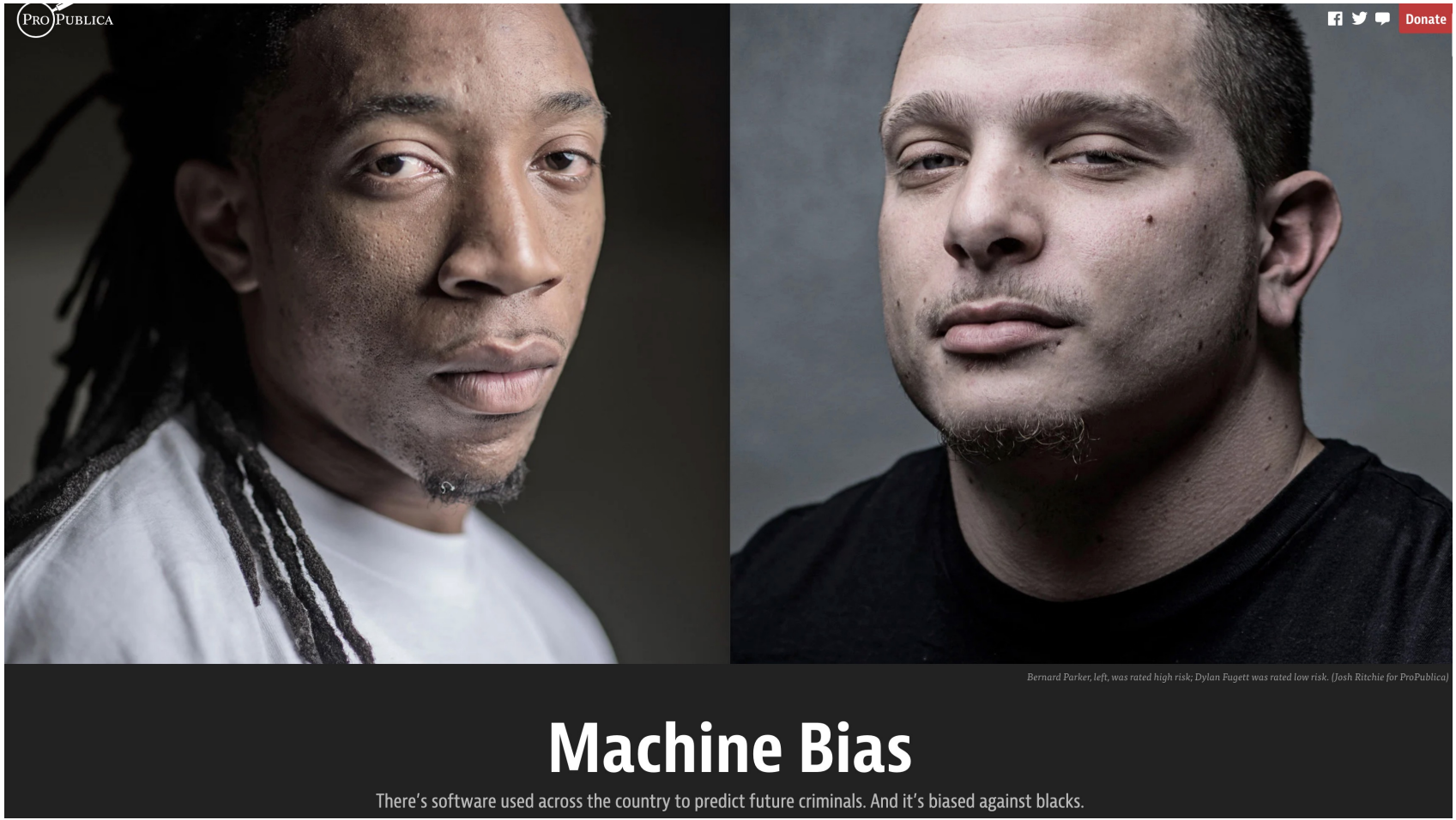
| | | | |
|---|---|---|---|
| ☐ Homicide | ☑ Weapons | ☑ Assault | ☐ Arson |
| ☐ Robbery | ☐ Burglary | ☐ Property/Larceny | ☐ Fraud |
| ☐ Drug Trafficking/Sales | ☐ Drug Possession/Use | ☐ DUI/CUIL | ☑ Other |
| ☐ Sex Offense with Force | ☐ Sex Offense w/o Force | | |

1. Do any current offenses involve family violence?
   ☑ No ☐ Yes
2. Which offense category represents the most serious current offense?
   ☐ Misdemeanor ☐ Non-violent Felony ☑ Violent Felony
3. Was this person on probation or parole at the time of the current offense?
   ☑ Probation ☐ Parole ☐ Both ☐ Neither
4. Based on the screener's observations, is this person a suspected or admitted gang member?
   ☐ No ☑ Yes
5. Number of pending charges or holds?
   ☑ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4+
6. Is the current top charge felony property or fraud?
   ☑ No ☐ Yes

**Criminal History**

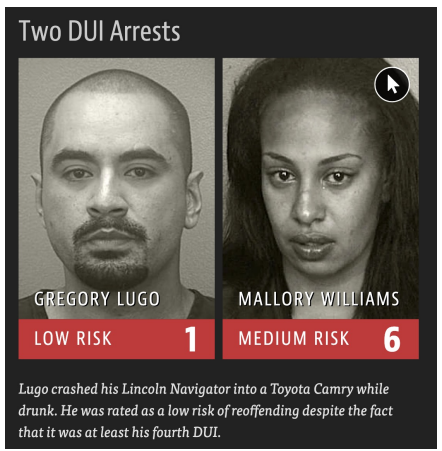
**Exclude the current case for these questions.**

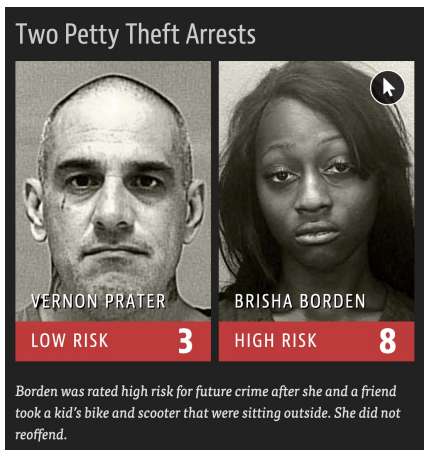
# Biases in COMPAS




https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing






*"In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.*

- *The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.*
- *White defendants were mislabeled as low risk more often than black defendants."*

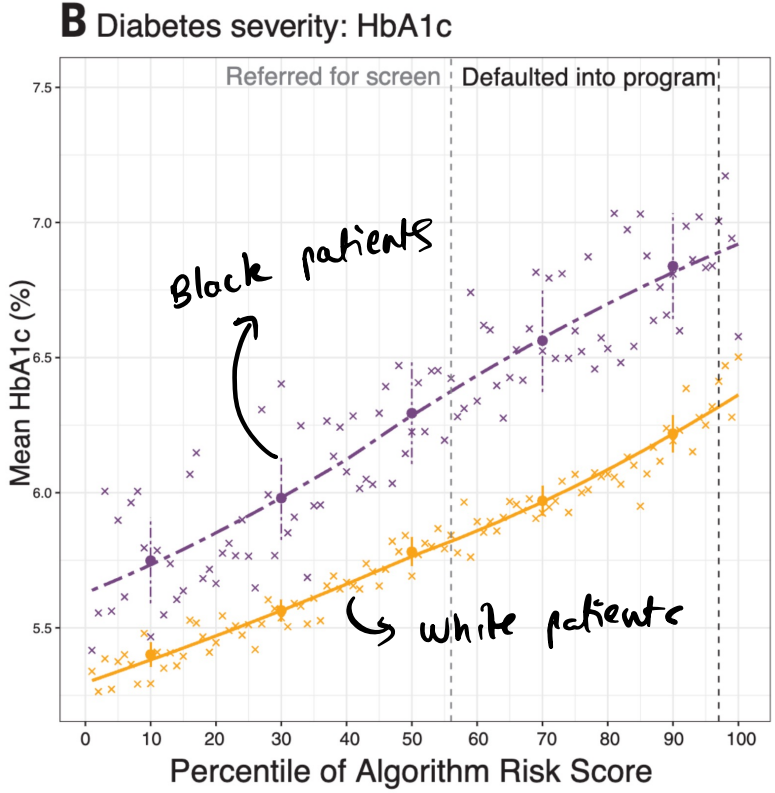
We will also see later that there are inherent tensions here: the COMPAS algorithm is biased in one way and unbiased in another, and it may be impossible to simultaneously be unbiased in both.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Bias in predictions: Predicting disease severity

Quoting from the paper:

- Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs.
- A widely used algorithm affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses.
- Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%.
- Bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means typically less money was spent on care for Black patients than for White patients.




Dissecting racial bias in an algorithm used to manage the health of populations, Obermeyer et al., Science 2019

# Some more instances of algorithmic decision making gone wrong

Aug 19, 2020 - Technology

## How an AI grading system ignited a national controversy in the U.K.

Bryan Walsh, author of Axios Future

Illustration: Eniola Odetunde/Axios

A huge controversy in the U.K. over an algorithm used to substitute for university-entrance exams highlights problems with the use of AI in the real world.

Link to article


Link to article

# Some more instances of algorithmic decision making gone wrong

The New York Times

## *There Is a Racial Divide in Speech-Recognition Systems, Researchers Say*

Technology from Amazon, Apple, Google, IBM and Microsoft misidentified 35 percent of words from people who were black. White people fared much better.

Give this article

Amazon's Echo device is one of many similar gadgets on the market. Researchers say there is a racial divide in the usefulness of speech recognition systems. Grant Hindsley for The New York Times

Link to article

MIT Technology Review

Featured Topics Newsletters Events Podcasts Sign in Subscribe

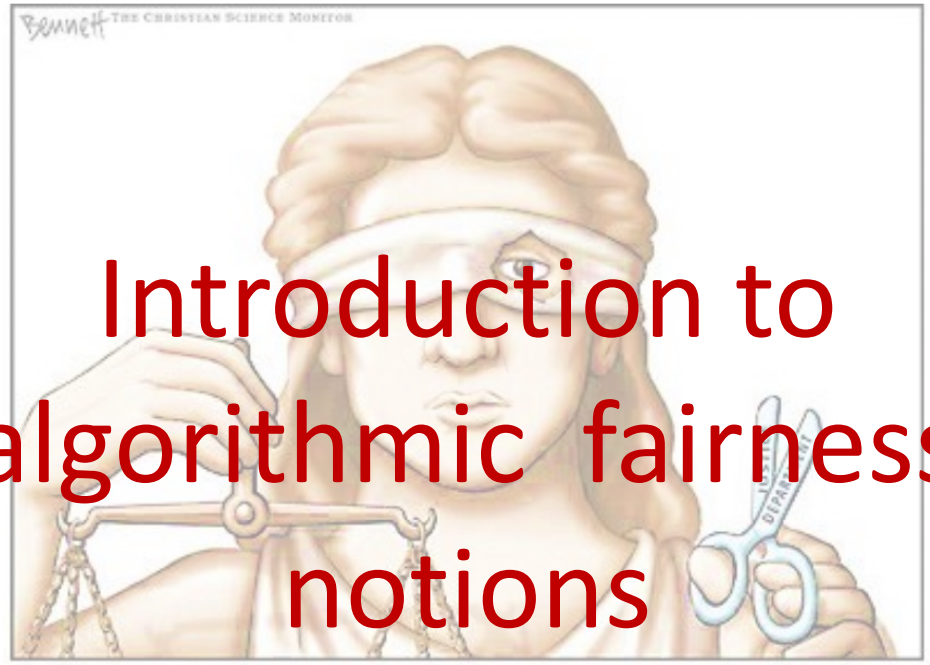
ARTIFICIAL INTELLIGENCE

## LinkedIn's job-matching AI was biased. The company's solution? More AI.

ZipRecruiter, CareerBuilder, LinkedIn—most of the world's biggest job search sites use AI to match people with job openings. But the algorithms don't always play fair.

By Sheridan Wall & Hilke Schellmann June 23, 2021

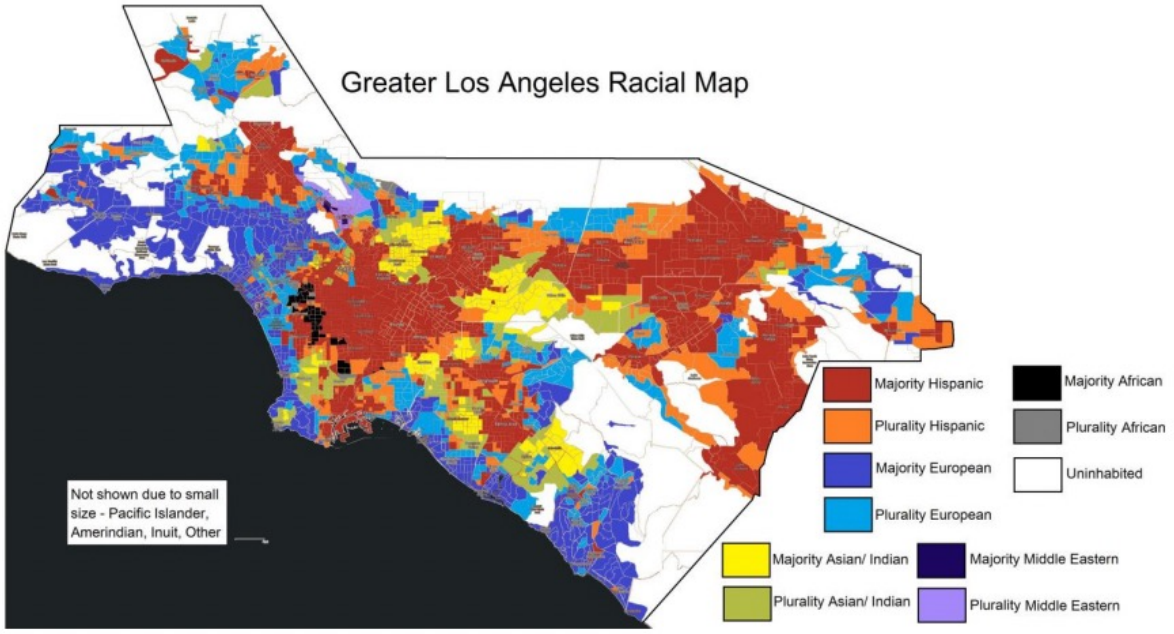
Link to article




# Introduction to algorithmic fairness notions

# How to obtain fair classifiers?

Observation: No fairness by just excluding sensitive attributes
Why? Sensitive attribute can often be reconstructed from other features




Zip code has a lot of information about race

# Ensuring fairness in classification: **Group & Individual fairness notions**

Two broad classes of fairness notions in classification:

**Individual fairness:** Algorithm treats **similar individuals similarly**

**Group fairness:** Algorithm is **"unbiased" on protected groups** (such as race, gender etc.)

# Individual fairness

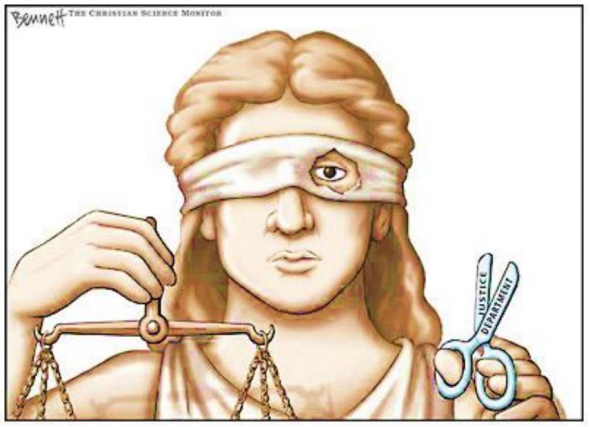
Define a ***metric*** $d(x, x')$ for the similarity between any two individuals $x$ and $x'$.

e.g.: $d(x, x') = \| x - x' \|_2$

If classifier predicts $p(x)$ as the probability of label being one for $x$, if

$$|p(x) - p(x')| \leq \mu \, d(x, x'),$$

then predictions of the classifier are individually fair with parameter $\mu$.





If these two individuals are similar, then their risk scores should be similar.

Fairness Through Awareness. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, Richard Zemel. 2011

# Group fairness

Group fairness notions require that the models predictions obey certain properties over protected groups (e.g. by race, gender).

Many different notions have been proposed

- Statistical parity
- Equalized odds
- Calibration across groups

# Statistical parity/Demographic parity

Binary classification setup (e.g. admitting a student to a degree program)

- Classifier $f$
- Datapoint $(x, y)$
- Sensitive attribute $a \in \{0,1\}$

Statistical parity (also known as demographic parity): $\Pr_x[f(x) = 1 \mid a = 1] = \Pr_x[f(x) = 1 \mid a = 0]$

In words: **Predictions are independent of sensitive attribute**

E.g., admit equal fraction of men or women into program

Can be too strong if labels and sensitive attribute are not independent.

E.g. if one demographic is more likely to be qualified than the other

# Equalized odds & Equality of opportunity

Same binary classification setup (e.g. admitting student to degree program)

- Classifier $f$
- Datapoint $(x, y)$
- Sensitive attribute $a \in \{0,1\}$

Just having 1st constraint (for $y = 1$) gives **equality of opportunity**

**Equalized odds**: Following 2 constraints are satisfied

1. $\Pr_{x,y}[f(x) = 1 \mid a = 1, y = 1] = \Pr_{x,y}[f(x) = 1 \mid a = 0, y = 1]$
2. $\Pr_{x,y}[f(x) = 0 \mid a = 1, y = 0] = \Pr_{x,y}[f(x) = 0 \mid a = 0, y = 0]$

*Recall* for class 1 = $\Pr_{x,y}[f(x) = 1 \mid y = 1]$

*FPR* = $\Pr_{x,y}[f(x) = 1 \mid y = 0]$

Equivalently:
**Recall** for $y = 1$ is the same for both groups (1st condition)
False positive rate (**FPR**) is the same for both groups (2nd condition)

Also equivalent to saying: Conditioned on label, prediction is independent of sensitive attribute

From *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

# Equalized odds

E.g. Professor Snape has to admit students to his Advanced Potions class.

100 students apply from Slytherin (80% are qualified)

| | Qualified | Unqualified |
|---|---|---|
| Accepted | 60 | 5 |
| Rejected | 20 | 15 |
| Total | 80 | 20 |

100 students apply from Gryffindor (40% are qualified)

| | Qualified | Unqualified |
|---|---|---|
| Accepted | 30 | 15 |
| Rejected | 10 | 45 |
| Total | 40 | 60 |

Is Prof. Snape fair based on

(i) statistical parity,

(ii) equalized odds?

Statistical parity :

$$\Pr_{x,y}[f(x) = 1 \mid a = 1] = \Pr_{x,y}[f(x) = 1 \mid a = 0]$$

Equalized odds:

$$\Pr_{x,y}[f(x) = 1 \mid a = 1, y = 1] = \Pr_{x,y}[f(x) = 1 \mid a = 0, y = 1]$$

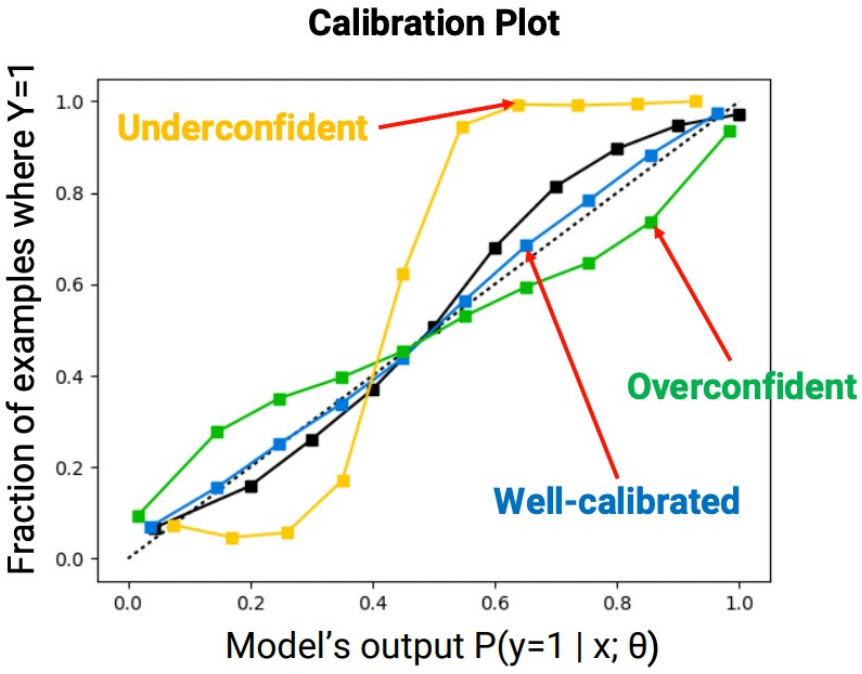
$$\Pr_{x,y}[f(x) = 0 \mid a = 1, y = 0] = \Pr_{x,y}[f(x) = 1 \mid a = 0, y = 0]$$

# Calibration

**Calibration**: A model $f$ for binary classification is calibrated if

$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha] = \alpha$$

Informally, this says that “predictions mean what they should”



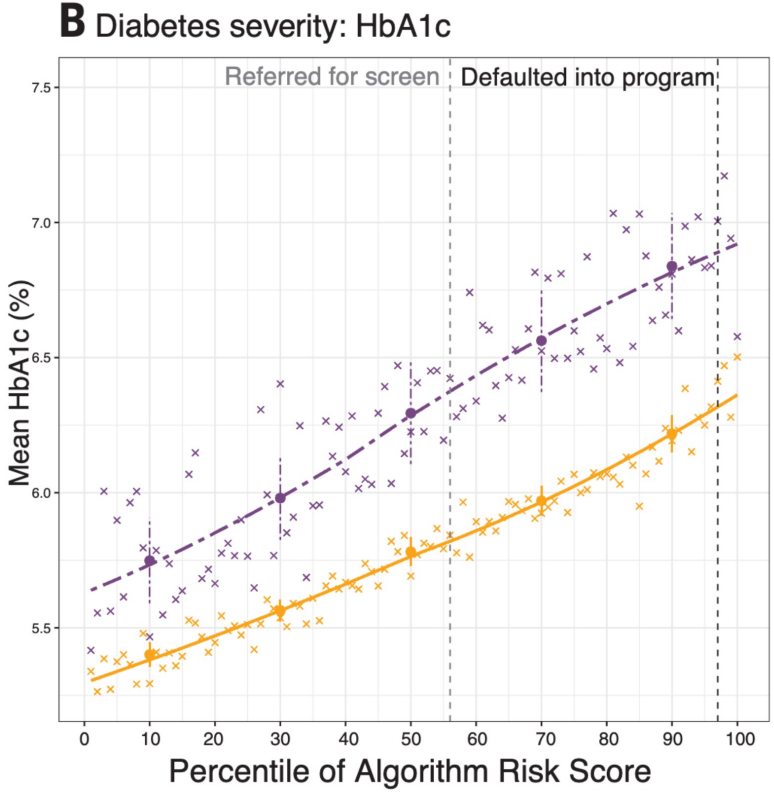

This is known as a ***reliability diagram***

# Calibration across groups

A model $f$ for binary classification is calibrated for groups defined by sensitive attribute $a$ if

$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, a = 1] = \alpha\,,$$
$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, a = 0] = \alpha.$$

Informally, this says that “predictions mean what they should for each group”



# Achieving these notions: Post-processing for statistical parity

Consider binary classification setup (e.g. admitting a student to a degree program)

- Predictor $p$ which predicts a score in $[0,1]$ (higher score => higher probability of label 1)
- Datapoint $(x, y)$
- Sensitive attribute $a \in \{0,1\}$

We want use $p$ to get classifier $f$ which maximizes accuracy but obeys statistical parity:

$$\Pr_x[f(x) = 1 \mid a = 1] = \Pr_x[f(x) = 1 \mid a = 0]$$

- What would $f$ be if we only wanted to maximize accuracy? (Suppose $p(x) = \Pr[y = 1|x]$)

- How to ensure statistical parity?

# Achieving these notions: Post-processing for equality of opportunity

Consider binary classification setup (e.g. admitting a student to a degree program)

- Predictor $p$ which predicts a score in $[0,1]$ (higher score => higher probability of label 1)
- Datapoint $(x, y)$
- Sensitive attribute $a \in \{0,1\}$

We want use $p$ to get classifier $f$ which maximizes accuracy but obeys equality of opportunity:

$$\Pr_{x,y}[f(x) = 1 \mid a = 1, y = 1] = \Pr_{x,y}[f(x) = 1 \mid a = 0, y = 1]$$

- How to obtain equality of opportunity?

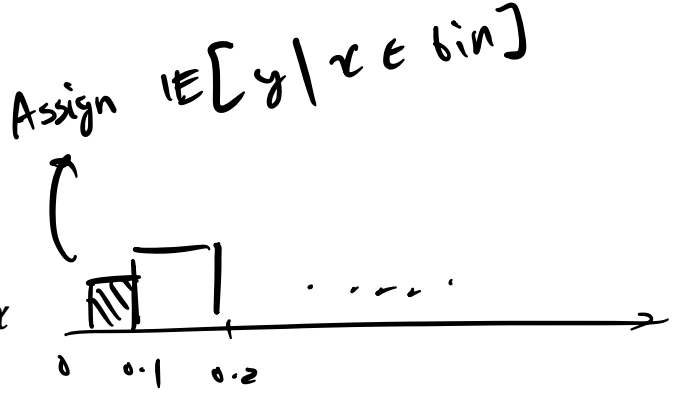
# Achieving these notions: Post-processing for calibration

Consider binary classification setup (e.g. admitting a student to a degree program)

- Predictor $p$ which predicts a score in $[0,1]$ (higher score => higher probability of label 1)
- Datapoint $(x, y)$
- Sensitive attribute $a \in \{0,1\}$

We want use $p$ to get classifier $f$ which is calibrated:

$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha] = \alpha$$




Basic idea: Suppose we have datapoints on which $p$ has the same score, say 0.3. Assign all of these datapoints to score $f(x) = \Pr[y = 1 \mid p(x) = 0.3]$

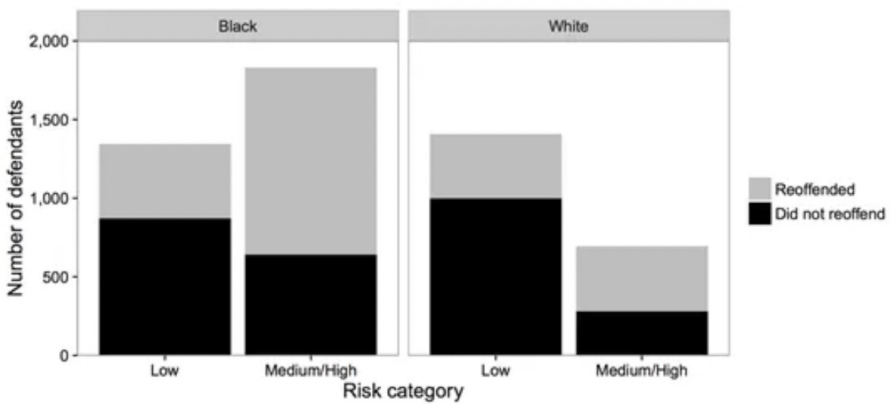
Histogram binning and Isotonic regression implement and generalize this.

# Group fairness notions: Can we satisfy them all?

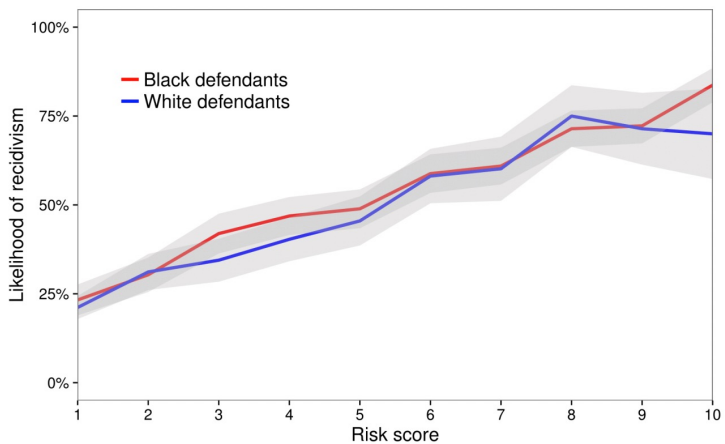
We saw three notions: statistical parity, equalized odds, calibration across groups
Can we always satisfy all of them together? ***No!***

In our example from Hogwarts, the model was fair in terms of equalized odds but unfair in terms of statistical parity. This tension between different notions arises in real data too.

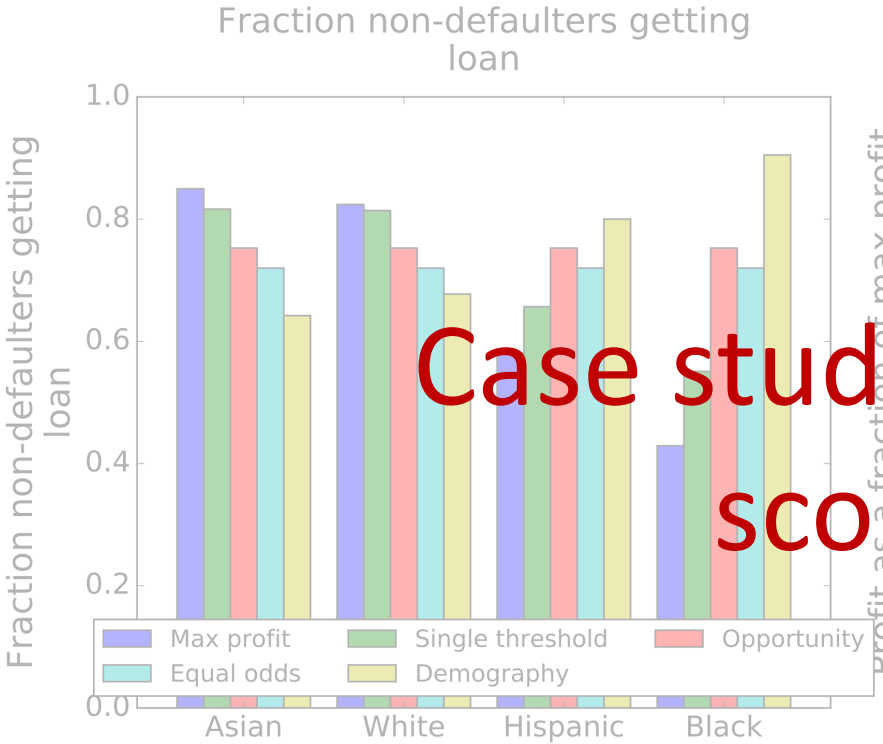
**COMPAS: Unfair** because black defendants who did not recommit crime are assigned higher score (i.e. does not obey equalized odds)

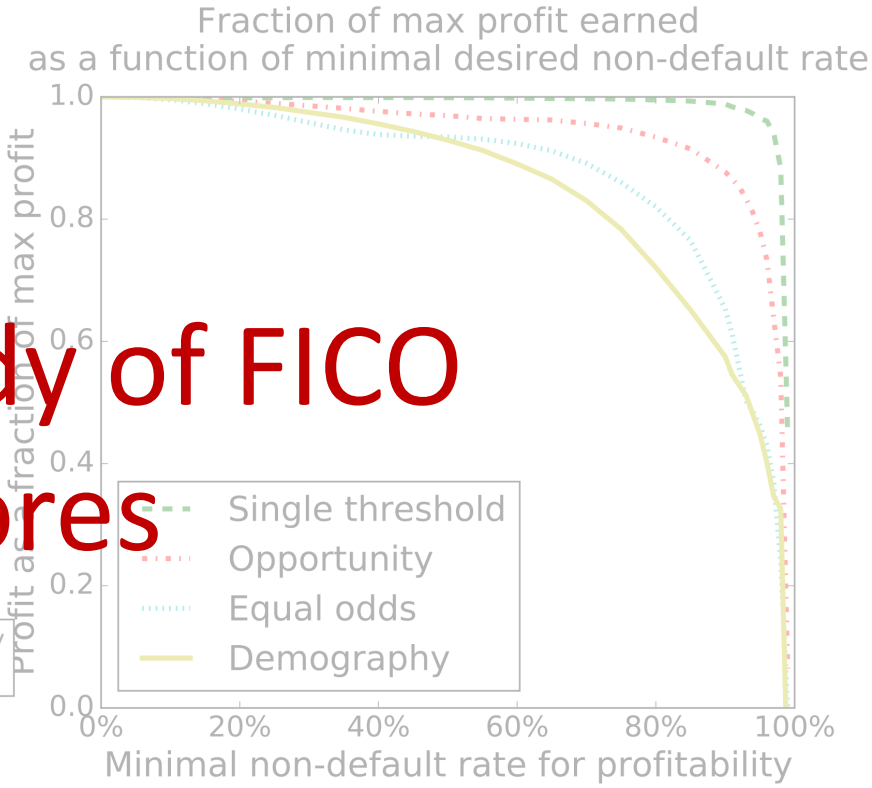



**COMPAS: Fair** because probability of recommitting crime is similar for a given risk score, for both groups (i.e. is calibrated)




https://medium.com/soal-food/what-makes-an-algorithm-fair-6ad64d75dd0c

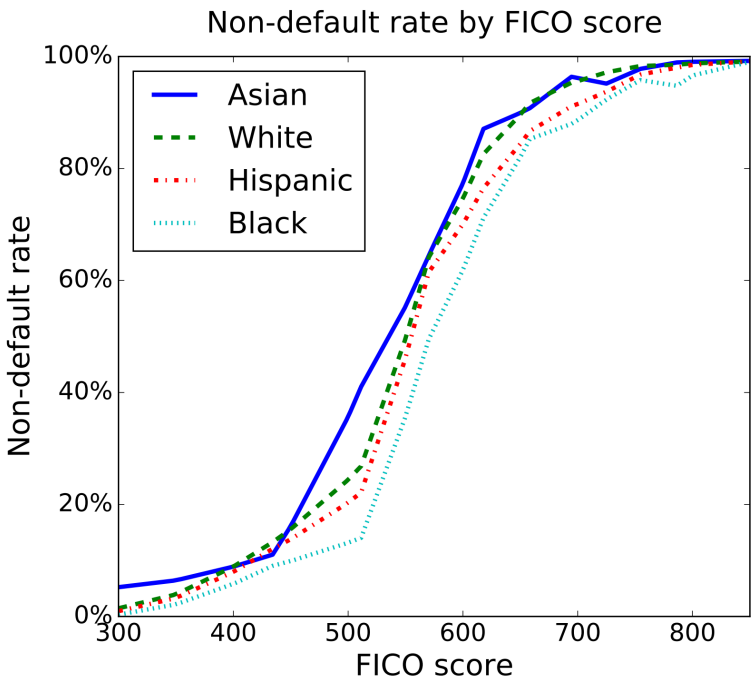



# Case study of FICO scores

# The input data: FICO scores and their distribution by race

- FICO scores based on 300k TransUnion scores from 2003
- Range from 300-850, aim to predict risk of defaulting on a loan
- 620 is common threshold for good loan rates. Corresponds to 82% non-default rate in the data



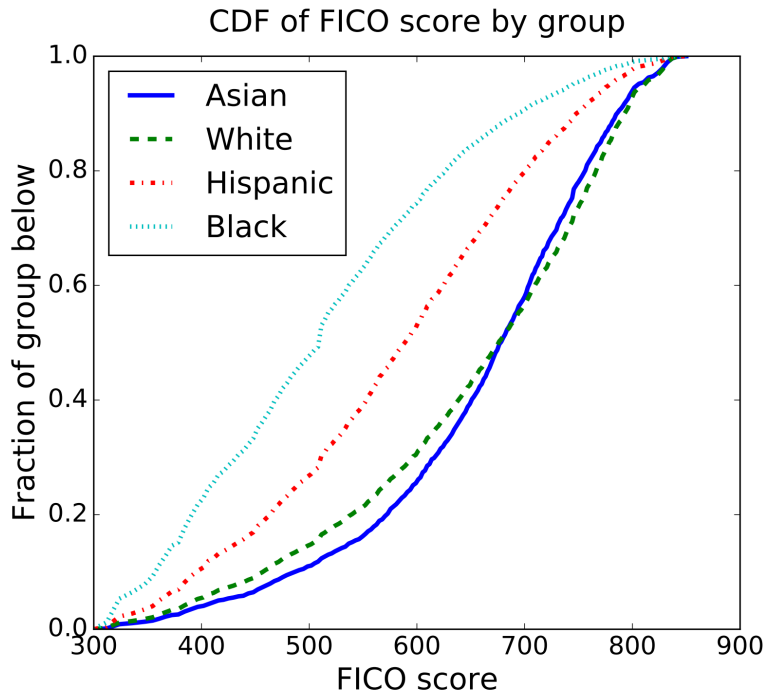
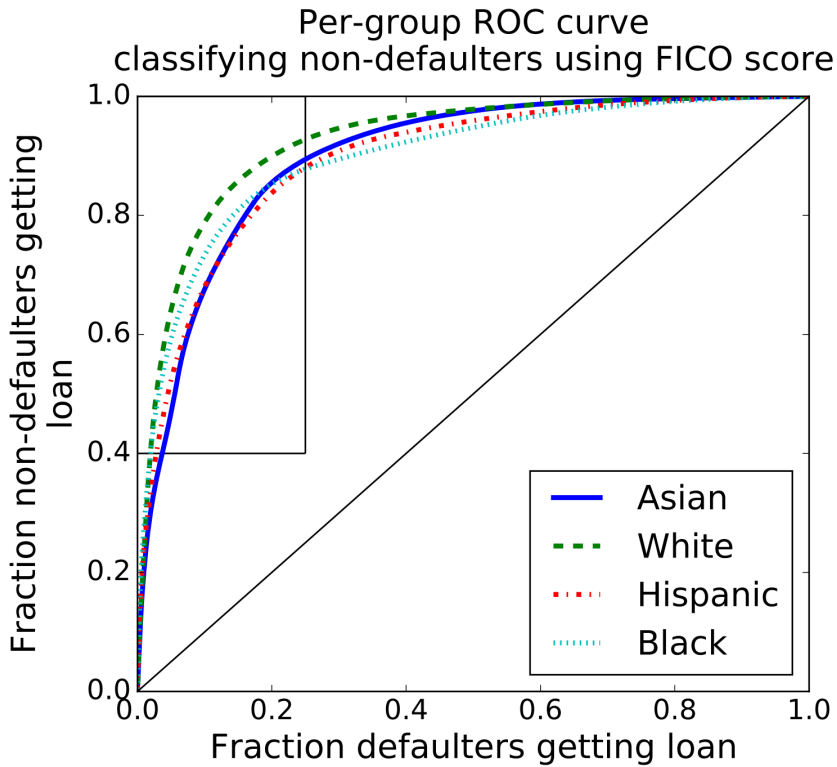




Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

# The decision-making rules

- **Max profit:** No fairness constraints, will pick for each group the threshold that maximizes profit. This is the score at which 82% of people in that group do not default.
- **Race blind:** Requires threshold to be same for each group. Hence will pick the single threshold at which 82% of people do not default overall.
- **Demographic parity:** Same as statistical parity, ensures fraction of group members that get loan is same across groups.
- **Equal opportunity:** Picks for each group a threshold such that fraction of non-defaulting group members that qualify for loans is the same.
- **Equalized odds:** Requires both fraction of non-defaulters that qualify for loans and fraction of defaulters that qualify for loans to be constant across groups.

# How accurate is the base classifier across groups?



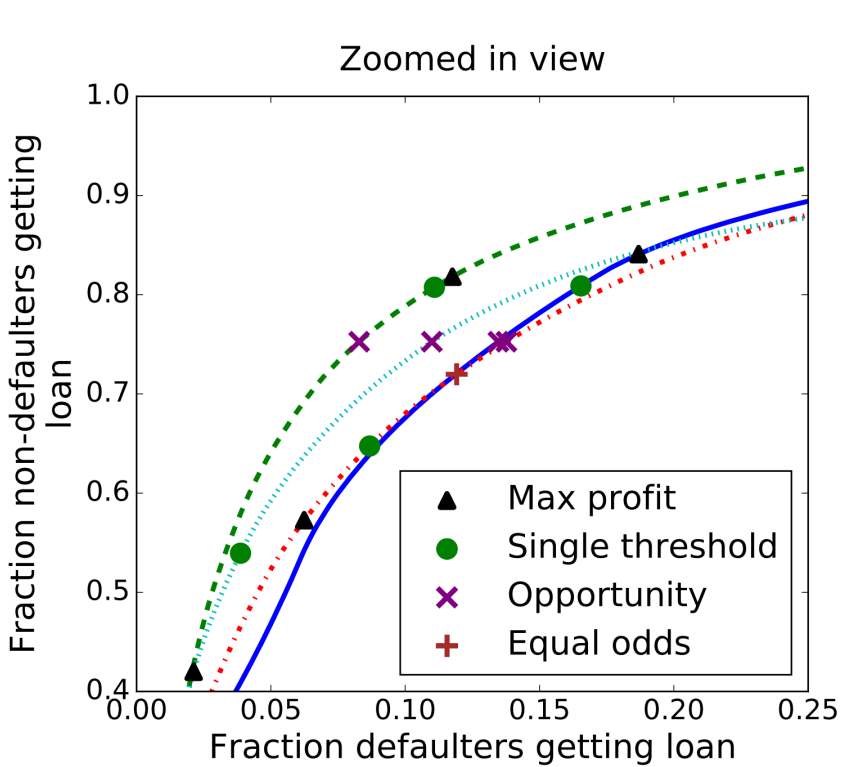
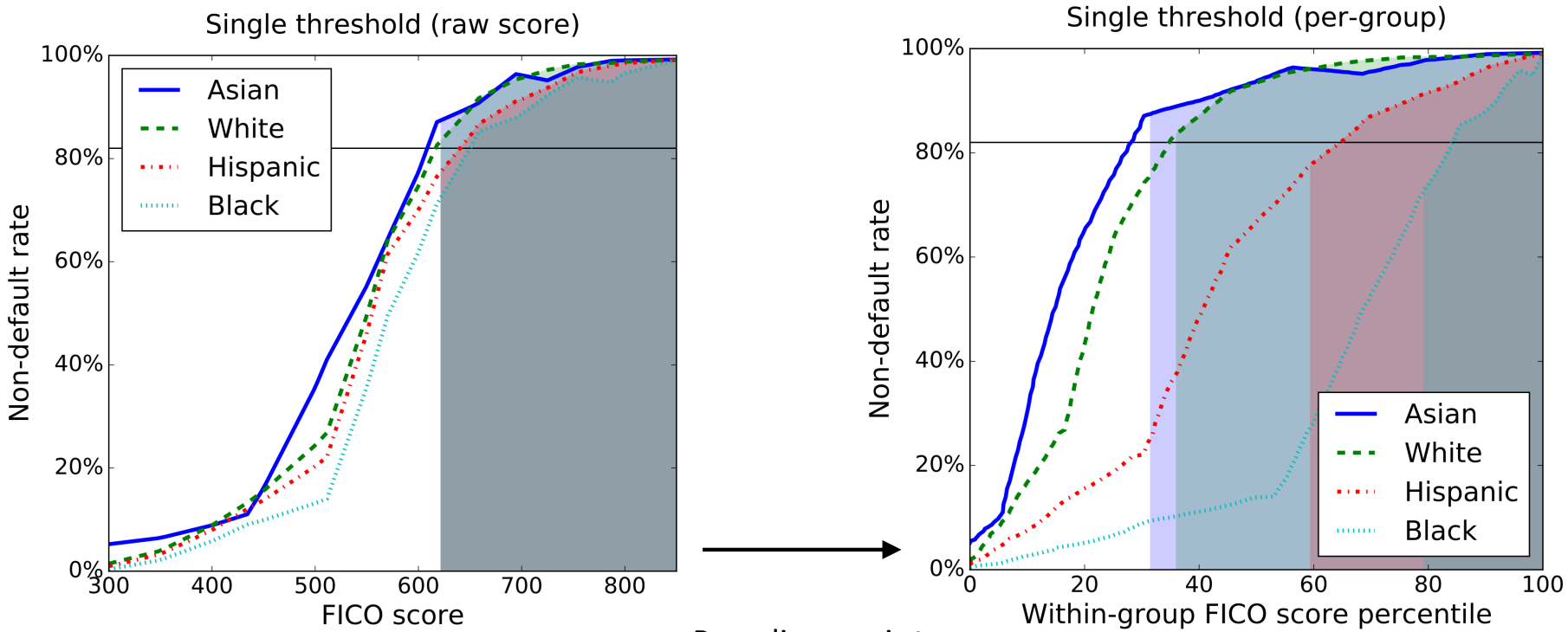




Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

# What happens when we pick a single threshold?

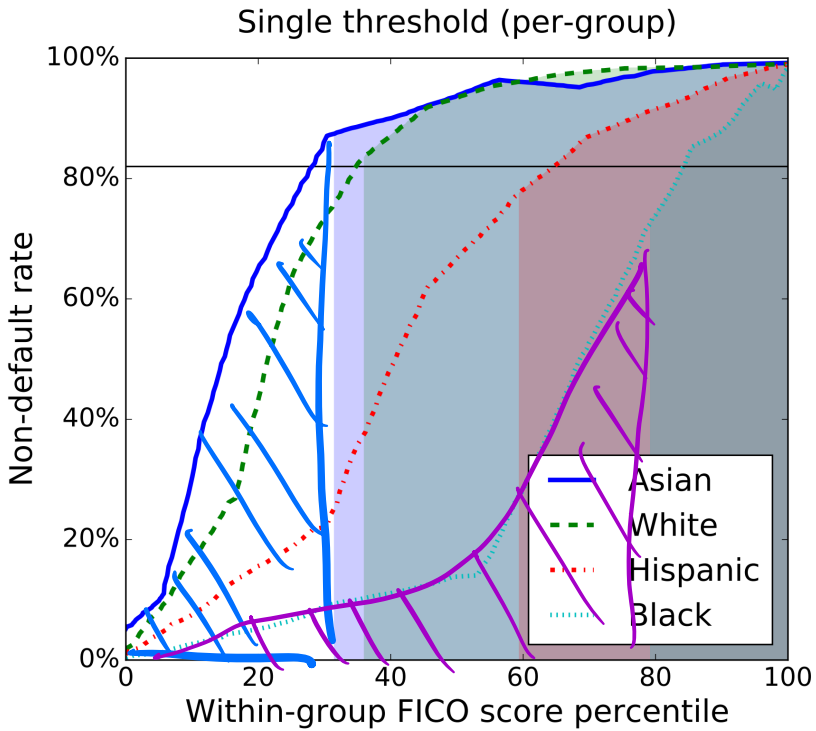



Rescaling x-axis to represent within group thresholds

Claim: $\Pr[f(x) = 1 \mid y = 1, a]$ is fraction of area under the curve that is shaded

Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

# Equality of opportunity as area under curve




Claim: $\Pr[f(x) = 1 \mid y = 1, a]$ is fraction of area under the curve that is shaded

Reason:

Consider any within-group percentile $t$.

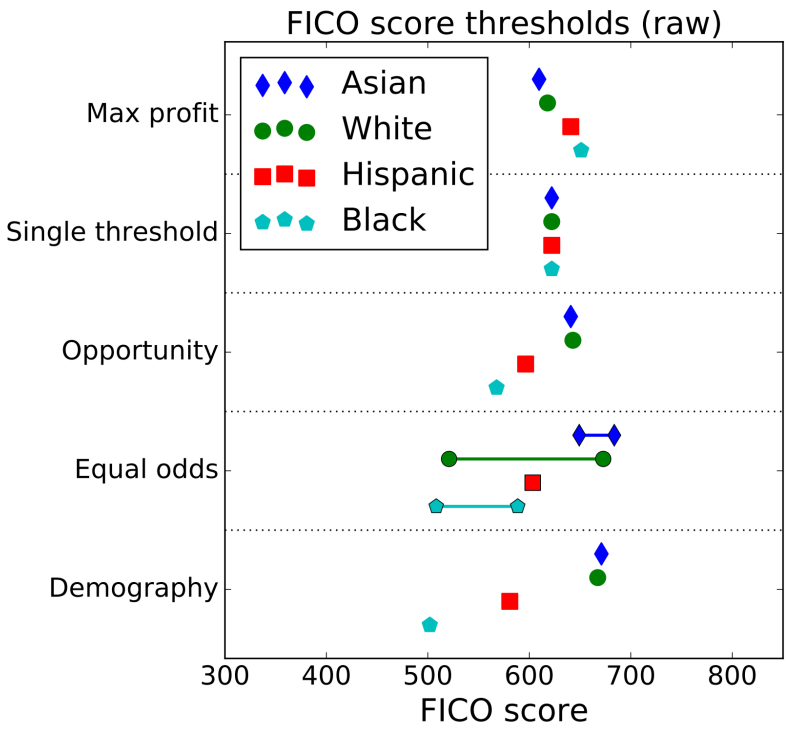
$\Pr[y = 1 \mid a] = \int_t \Pr[y = 1 \mid t, a\,]\ dt$ = area under curve

$\Pr[f(x) = 1, y = 1 \mid a] = \int_t 1\,(\,t > \text{threshold})\ \Pr[y = 1 \mid t, a\,]\ dt$

= area under curve which is shaded

Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

# What are thresholds for different rules?



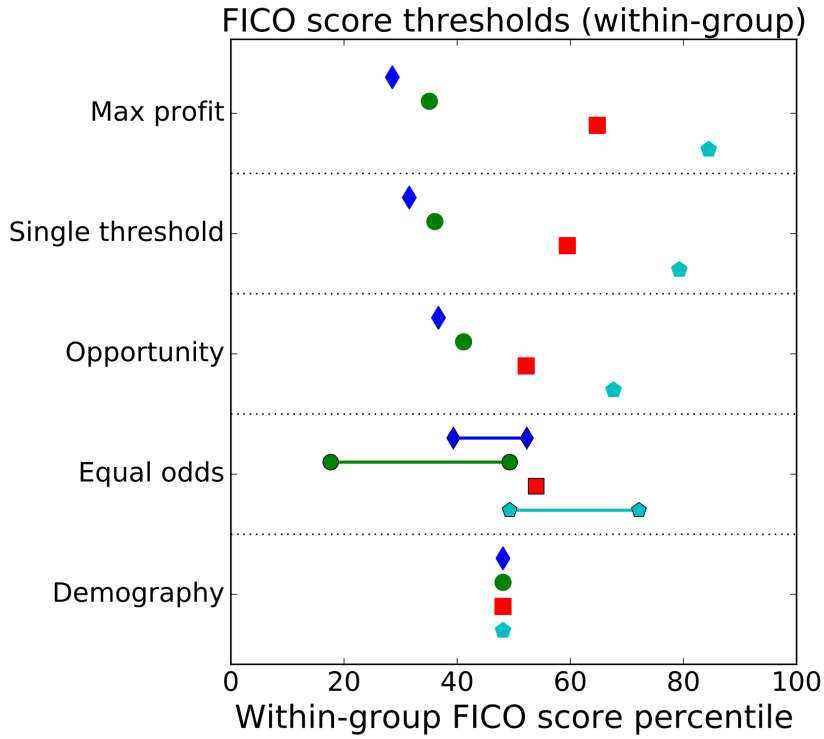
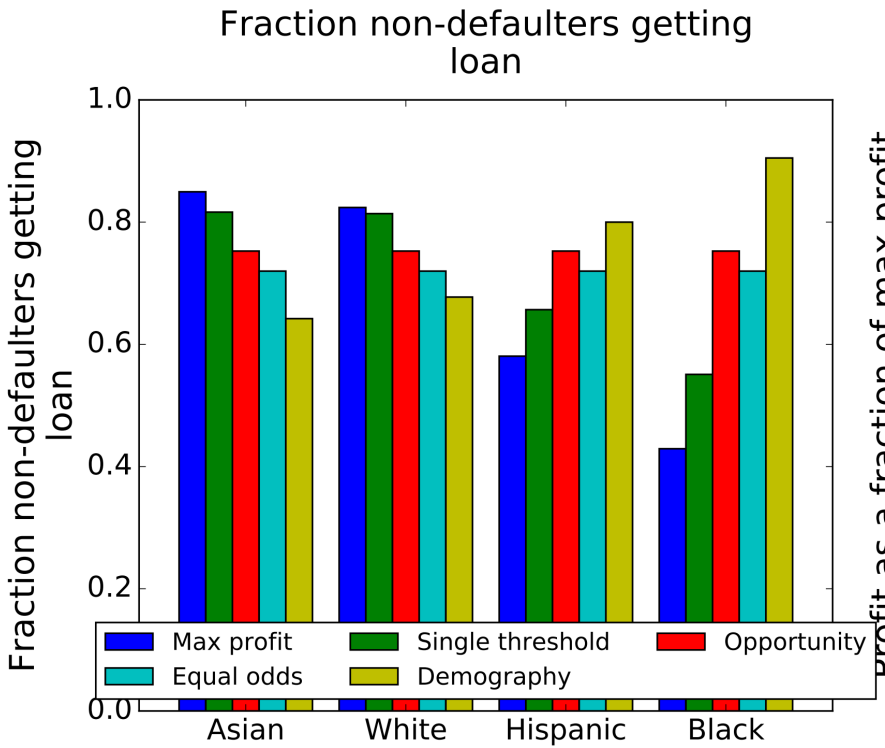




Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

# What fairness and utility do different rules obtain?



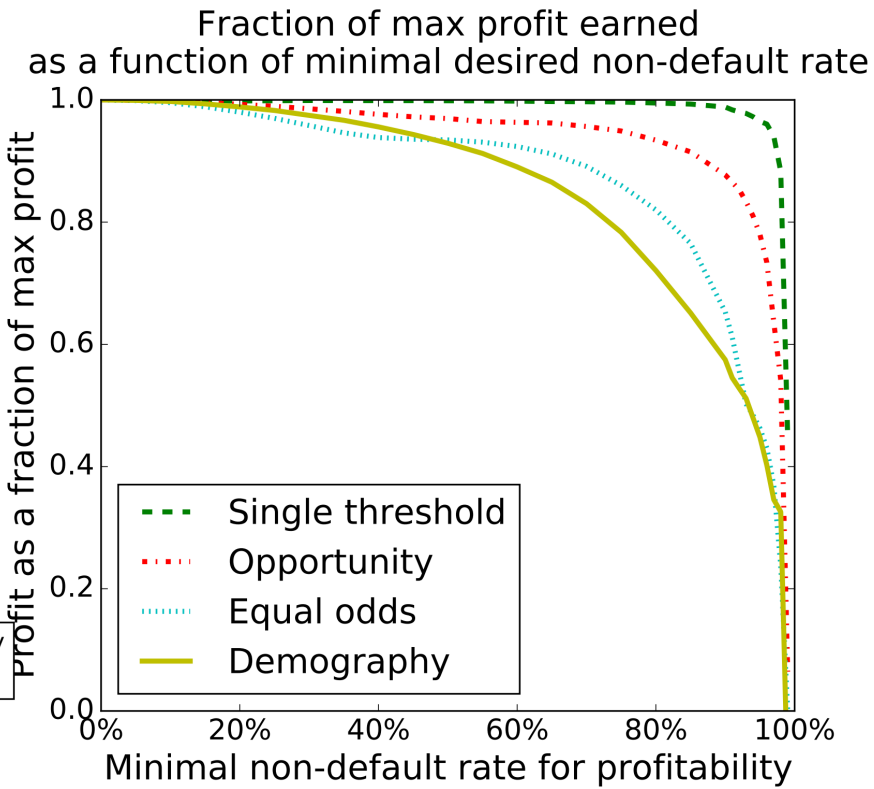




Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16